DEVCOM ANALYSIS CENTER

DEVCOM DAC-TN-2023-008
July 2023

The Use of Adjacent Video Frames to Increase Convolutional Neural Network Classification Robustness in Stressed Environments

by Patrick Debroux

DISCLAIMER

The findings in this report are not to be construed as an official Department of the Army position unless so specified by other official documentation.

WARNING

Information and data contained in this document are based on the input available at the time of preparation.

TRADE NAMES

The use of trade names in this report does not constitute an official endorsement or approval of the use of such commercial hardware or software. The report may not be cited for purposes of advertisement.

DEVCOM DAC-TN-2023-008
July 2023

The Use of Adjacent Video Frames to Increase Convolutional Neural Network Classification Robustness in Stressed Environments

by Patrick Debroux
DEVCOM Analysis Center

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
July 2023	Technical Note	1 October 2022	30 May 2023

4. TITLE AND SUBTITLE
The Use of Adjacent Video Frames to Increase Convolutional Neural Network Classification Robustness in Stressed Environments

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Patrick Debroux

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Director DEVCOM Analysis Center ATTN: FCDD-DAE-E White Sands Missile Range, NM 88002	DEVCOM DAC-TN-2023-008

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The study objective was to use adjacent video frames to increase the robustness of convolutional neural network (CNN) classifiers to stressed targets. We identified and downloaded video clips of targets that moderately change their aspect angle. Video clips of military vehicle target classes were previously used to fine-tune pretrained CNNs through transfer learning. We obtained the frame series from these video clips, and the target in each frame was stressed with different coherent stresses. Instead of relying on the classification of individual frame images, we used different running averages and running products on class probabilities of the classifiers to increase classification robustness to the applied stresses as the aspect angle of the target to the sensor changed. Our results showed modest changes in classifier robustness when we applied moving average/product filters to the output class probabilities. This robustness increase was most pronounced when a small number of elements were averaged, and it regained stability (at an increased robustness) as the filters we applied increased in element size.

15. SUBJECT TERMS
convolutional neural network, CNN classifiers, stressed target environments, classification robustness, class probability moving average filters

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	27

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Patrick Debroux	(575) 678-5238

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

The use of artificial intelligence classifiers in military systems is somewhat different from those used in commercial industries due to differences in the images used to train the convolutional neural networks (CNNs) and the images the CNN will try to classify. These differences, which can be intentional or unintentional, make even a well-trained classifier eventually fail as the target image becomes unrecognizable.

It is our goal to make these CNN classification systems more robust to target stresses to extend their use in battlefield environments where targets are intentionally hidden and altered, sensors are obscured, and hostile actors intentionally interfere with sensor images. CNN model classification effectiveness and robustness have largely been optimized, and the state-of-the-art CNN is presently on a plateau.

CNN classifiers must be trained to recognize targets to classify them. Because of the possible variability of the stresses applied to targets, it is difficult and impractical to train military CNN classifiers on every possible type and intensity of target stress found on the battlefield; therefore, other ways to extend the robustness of classifiers to battlefield stresses must be found. Research is being conducted on classifier fusion[1] and sensor fusion for classification.[2] Researchers have previously constructed CNN architectures to use on the temporal aspect of video clips to train CNNs and classify objects directly from video, but much of that work has been spent on increasing the efficiency of the algorithmic processes, and after implementation, researchers found that although training the CNNs from video became more efficient, the classification did not significantly improve.[3]

In this report, we use multiple sensor images to extend the range of classification success of CNN classifiers when the classification targets become highly stressed. We used adjacent frames of sensor video, and instead of classifying the target in each frame, we added the class probability of the classifier in a running average before choosing the class with the largest probability as the correct answer. We believe this will increase the chance of successful classification in stressed environments. Since the probability of classification ranges from 0 to1 for each class, a moving product filter can also be used to increase the success of classification. The number of elements in the moving average and moving product filters can also be changed to detect their optimal lengths.

We hypothesize that as adjacent frames are used, especially when the target azimuth angle with respect to the sensor gradually changes, the relative stress locations over the targets may change, thus revealing new features used by the classifiers to correctly classify the target.

This idea parallels the radar signal stacking practice, where radar returns are added together to increase the response of the target and to reduce system noise. When we apply this to the classification of adjacent video frames, and if we use moving averages and moving products in object classification in stressed environments, then we may be able to develop more robust classification.

2. METHODS

To test this hypothesis, we found video clips of military land vehicles and downloaded them from the web. The vehicle video clips we chose coincided with the military vehicle training database previously developed for CNN classification analysis in stressed environments.[4,5] We specifically sought video clips that showed the target with slowly changing viewing angles. The six classes developed with transfer learning in pretrained CNNs are as follows:

- Abrams Tank
- Bradley Infantry Fighting Vehicle
- High Mobility Multipurpose Wheeled Vehicle (HMMWV)
- Cougar Mine-Resistant Ambush-Protected (MRAP) 6×6
- Interim Armored Vehicle Stryker
- T-72 Tank

Once the video clips were downloaded, we edited them to extract those segments that showed the unobstructed target vehicle isolated from other vehicles, with enough color contrast not to blend significantly into the background. This color contrast was needed to isolate the target from its background quasi-automatically. Some of the video clips we downloaded from the web yielded several usable video segments in which the isolated target vehicle stayed within relatively similar backgrounds.

The video segments were then separated into frames, and each frame became an image containing the target to be classified by the CNN. Because a video is normally recorded at 30 frames per second, adjacent frames are not expected to greatly vary from each other.

We chose a prototypic frame that represented the segment's target and background from the frame series, and we used the MATLAB k-mean clustering function (imsegkmeans.m) in the Image Processing Toolbox to manually separate the frame image into three parts: the target, the background that makes up the sky, and the rest of the background that makes up the ground, foliage, buildings, and so on (Figure 1). The prototypic frame was usually chosen as one of the first few frames of each segment unless visual inspection revealed an abnormality when compared to adjacent frames.

Figure 1. Example of image processing to define target and background masks

With the assumption that adjacent frames have similar target and background hues (again, video segments are typically recorded at 30 frames per second), the cluster labels (target, earth, and sky) defined in the prototypic frame were automatically applied to the k-mean clusters defined in the rest of the frames of the video segment. This automatization process allowed us to label the frames of the video segments relatively quickly. Although this procedure was not perfect, it was efficient for most of the frames we considered.

We performed a manual quality check on the frame images that were reduced to their labeled logical masks. The mask images that did not follow the prototypic mask image scheme were rejected. We found that the segment choice defined for the prototypic frames drifted away from the original mask choice for video frames that were very distant from the prototypic frame.

Throughout this thinning process, we kept the video segments presented in Table 1. Table 1 also shows the number of usable frames in each segment. The frames are identified as usable if they have the same k-cluster labels as that defined in the prototypical frame.

Table 1. Video segments and the number of usable frames they contain

Clip name	No. of usable images
abrams_1_a	31
abrams_6_a	154
bradley_2_a	88
bradley_2_d	50
bradley_5_a	53
bradley_7_a	123
hmmwv_3_a	348
hmmwv_5_a	28
mrap_2_a	20
stryker_1_a	8
stryker_2_a	73
stryker_2_b	32
stryker_4_a	39
stryker_4_b	51
stryker_6_a	49
stryker_6_c	53
stryker_6_d	67

Now that the targets masks are labeled in the frames of the different video segments, the six different stresses previously developed[4] are applied incrementally to all the frames in the video segments. Incrementally applied synthetic target stress types are as follows:

- Target Camouflage
- Disruptive Coloration 1 (vertical)
- Disruptive Coloration 2 (horizontal)
- Target Distortion
- Target Hull Defilade
- Image Obscuration

Examples of incremental application of the six target stresses are presented in Figures 2–7, in 10% increments.

Figure 2. Generic camouflage applied to a target image in 10% increments

Figure 3. Vertical disruptive coloration applied to target in 10% contrast increments

Figure 4. Horizontal disruptive coloration applied to target in 10% contrast increments

Figure 5. Pixel and edge dilation used to mimic target distortion

Figure 6. Simulated obscuration by random pixel disruption in 10% increments

Figure 7. Target hull defilade simulation increased in 10% increments

Because the target similarity of adjacent frames is going to be used in this analysis, some improvements to the algorithm that produces the camouflage stress on the images were made. The camouflage stresses have to be “coherent” between frames, meaning that the camouflage patches have to appear at the same place on the target on each successive frame image, instead of being randomly placed in every frame.

We developed coherence by defining and storing the relative distance of each camouflage patch seed (these are the single pixels around which the camouflage patches will be developed) to the most top left pixel of the target mask. By saving these relative distances between adjacent frames and reapplying patch seeds relative to the upper-left corner of the target mask in the next frame, the relative location of the patches on the target is approximately preserved in adjacent frames, assuming the viewing angle of the target does not change too drastically. Figure 8 is a collection of adjacent frames and shows the location of the camouflage patches stay in approximately the same place over the target.

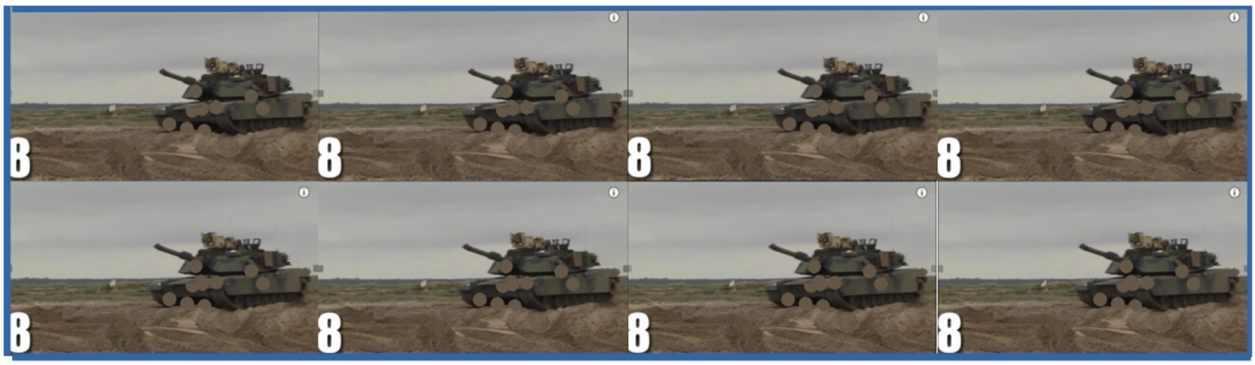

Figure 8. Series of adjacent frame images showing the coherent placement of camouflage patches

Making the camouflage stress coherent in adjacent video frames is important for both the reality of the stress and the quality of the analysis because the patches may or may not cover a local target geometry that the CNN uses to classify the object.

With all the image frames progressively stressed with different synthetic stresses, the following eight pretrained CNN models downloaded as MATLAB Deep Learning Toolbox Add-ons can be used to determine classification success of progressively increasing stressed images:

- Squeezenet
- Googlenet
- Resnet18
- MobilenetV2
- Resnet50
- Resnet101
- InceptionV3
- InceptionResnetV2

The CNNs were fine-tuned using transfer learning to the six classes using 7500 unstressed training images per class. Since these images were downloaded from the web, or obtained from downloaded video clips, they are images of vehicles in action and therefore may be obscured, stressed, or in complicated backgrounds. These “natural” stresses in the training images are not the same as those progressively applied to the test images, but they may play a small, unknown role in this analysis.

The images of each video segment were classified, and the classification robustness in stressed environments were calculated in two ways:

- The first is the conventional method of comparing the classification of each image against its class label to calculate the classification success level of the

CNNs as the stresses increase in the images. In this method, the classification of each image is simply the class that has the highest probability for that image.

- The second method is to use a running average filter over adjacent frames in a video sequence before calculating the class with the highest probability to compare against the video segment label. The running average elements are variable and vary from 1 to 20. In this analysis, the filter sizes are squeezed down at the beginning and end of the data sequence, or when the data set is smaller than the size of the filter. Because the probability ranges from 0 to 1, a running product filter can also be applied in a similar manner.

A comparison of the results of the two methods will either prove or disprove the hypothesis that “stacking” the class probability of classification before classification will increase the robustness of classifiers in stressed environments.

The classification success results are arranged in a multidimensional matrix, with dimensions of 8 × 17 × 6 × 20 × 21. These dimensions are defined by the 8 CNNs used for classification, the 17 video segments by the 6 stresses progressively applied to the images, then by the number of elements in the running average/product, and finally by the 5% increments of the applied stress.

The classification success analysis defined by method 1 is a specialization of this matrix with the running average/product set to 1. Using this matrix specialization between the two methods, our goal is to compare the classifier success versus image stress calculated with a moving average/product filter size of 1 compared to filters with more elements.

3. RESULTS

The results of this study are the comparisons of classification success curves for different sizes of running average/running product filters. Because of the large dimensionality of input variables, we need to provide an average for the parameters that are not immediately being considered.

The first dimension we averaged was the results over the various CNN models. The CNN models that were downloaded are all award-winning, pretrained models that performed equally well here. As a result, a comparison of CNN model results was not done. After averaging the CNN model results, we had a 17 × 6 × 20 × 21 matrix dimension.

This leaves the first two dimensions to be averaged out individually to plot the remaining curves for the dimension that has not been averaged. The video segments are averaged out, which leaves a 6 × 20 × 21 matrix dimension. Figure 9 shows six plots of classification success versus incremental target stress corresponding to the six stresses. Each plot shows a family of curves for different-sized running average/product filters run over the class probabilities before classification.

Figure 10 shows the same plots as Figure 9, but the family of curves in each plot is the classification success for different-sized running product filters used over the class probabilities before classification. Because class probabilities equal less than 1, no normalization is needed before plotting.

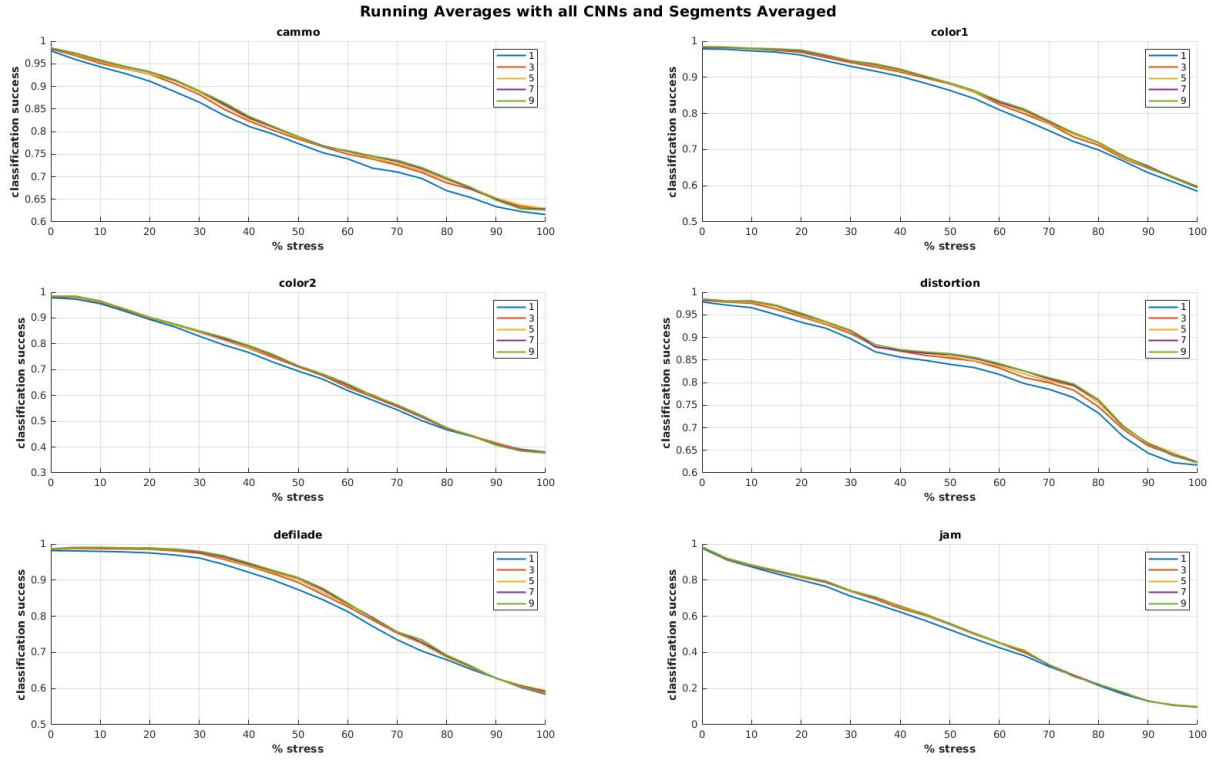

Figure 9. Classification success vs. target stress for different stresses. The family of curves are for different length moving average filters.

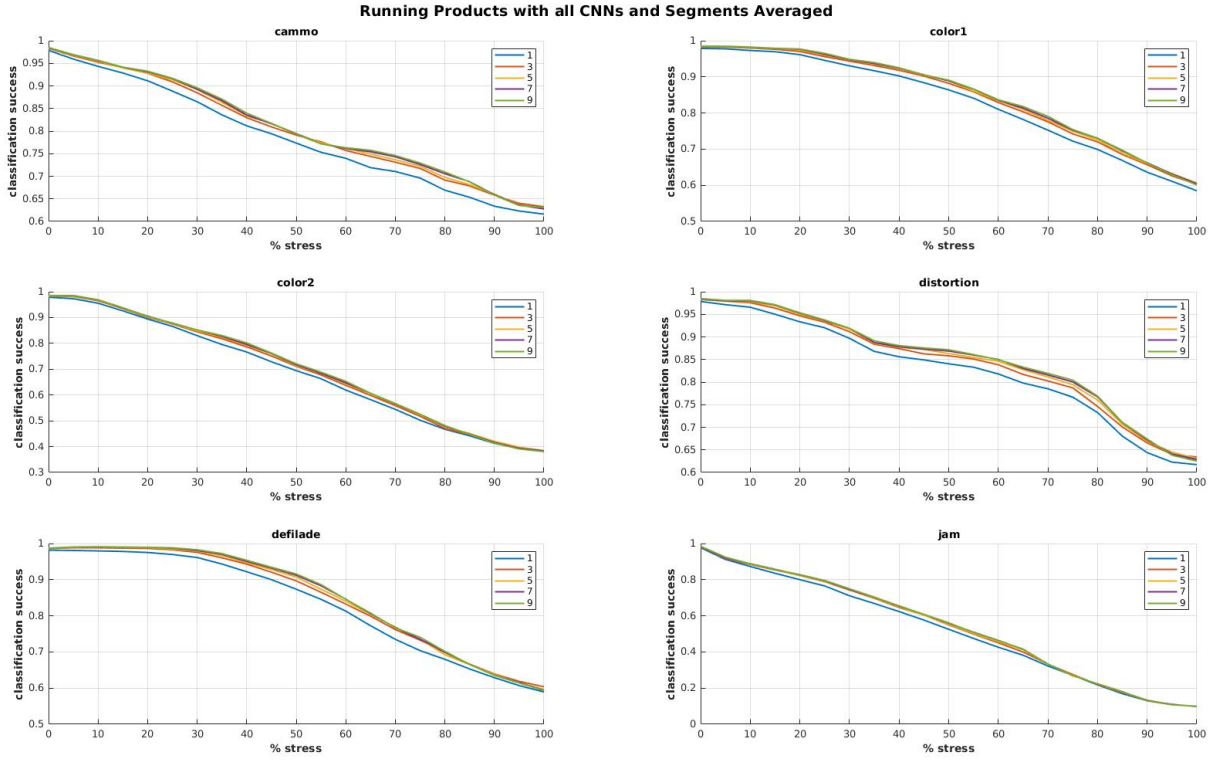

Figure 10. Classification success vs. target stress for different stresses. The family of curves are for different length moving product filters.

The corresponding family of curves in Figures 9 and 10 show that the moving average and product filters essentially yield the same results. We see a slight increase in robustness as the elements of the running average/product filters increase, but they soon stabilize at a filter length of seven elements. Disruptive coloration and the obscuration (jamming) stresses are helped the least by using averaging/multiplying the class probabilities over several images before classifying an image. The lack of robustness enhancement for the disruptive coloration stresses is believed to be caused by the constancy of the variable-width striping over the target and using increased contrast of the striping to incrementally increase the target stress. The CNN models did not greatly increase robustness with the class probability averaging/product for the obscuration stress because of the homogeneity of the stress over the whole image. A slightly different target viewing angle does not seem to help robustness for this kind of stress.

Next, the matrix is averaged over the stresses to look at the family of curves for different video segments. Figure 11 shows 17 plots of classification success versus incrementally averaged target stress for the different video segments.

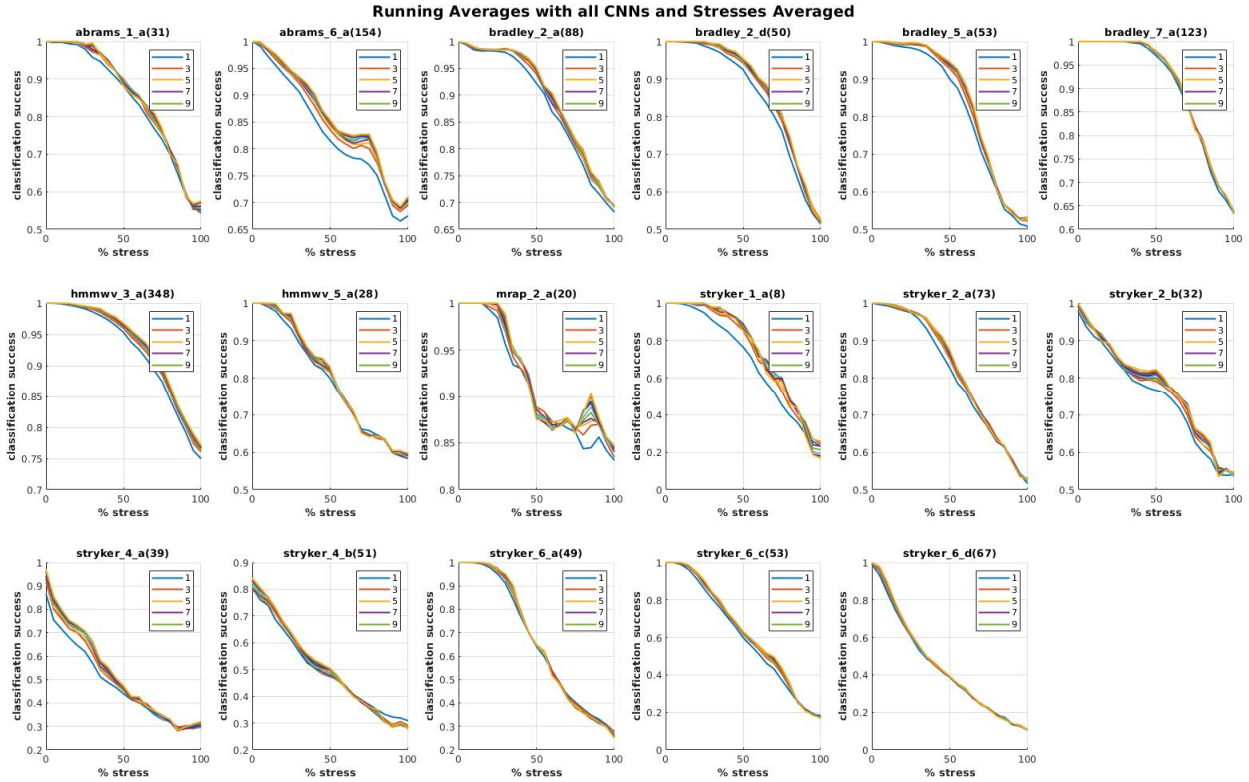

Figure 11. Classification success vs. target stress for different video segments. The family of curves are for different length moving average filters.

Figure 12 shows the classification success plots of various video segments with averaged target stresses. This time, the family of curves shows moving product filters of a different size.

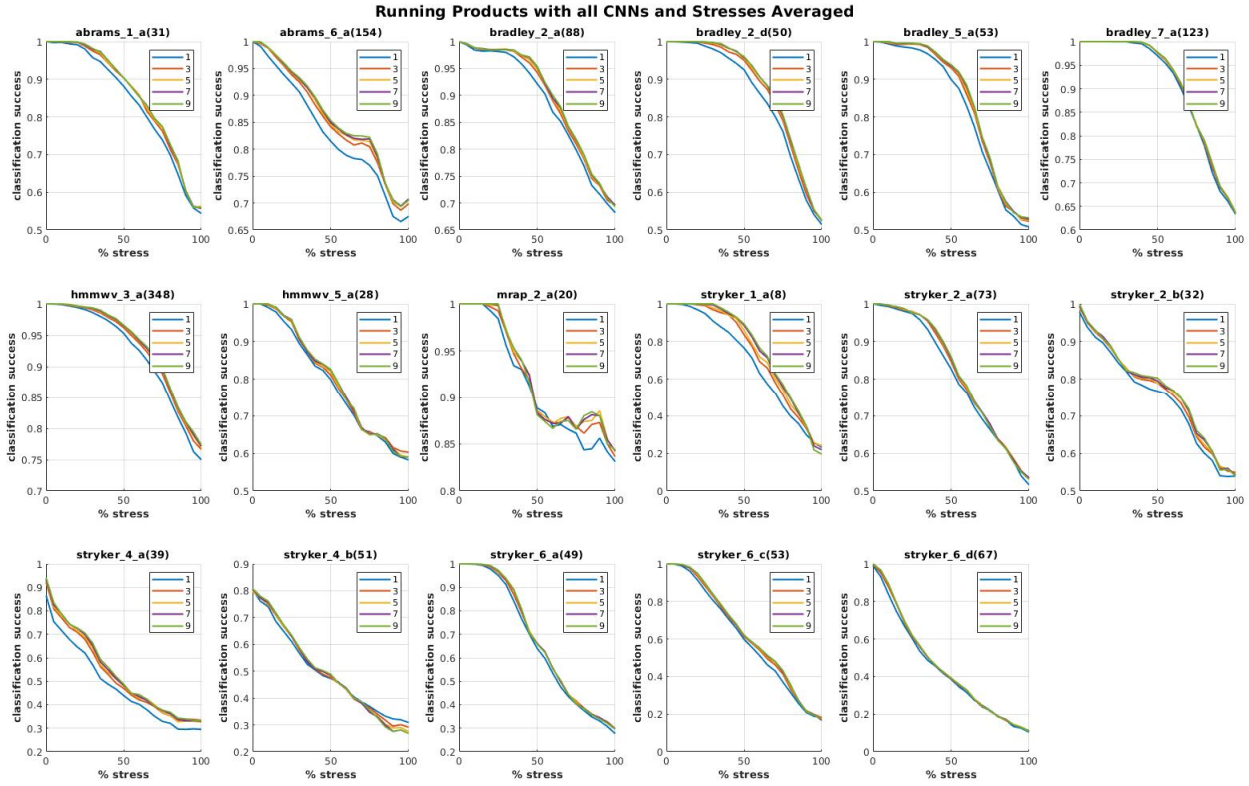

Figure 12. Classification success vs. target stress for different video segments. The family of curves are for different length moving product filters.

There is some variation in the robustness increase because of the running average and running product filters. This can be attributed to a lack of change in the viewing angle in adjacent frame images, which does not add additional information in the running averages/products of adjacent frames.

Finally, the last two plots show the family of curves of different filter lengths when both the video segment and the stress-type dimensions are averaged, leaving a 20 × 21 matrix dimension.

Figure 13 shows a family of curves for different length running averages when both the stress types and video segments are averaged. Figure 13 also shows a small increase in robustness when a running average filter is applied to class probabilities before the maximum probability is chosen to identify the target.

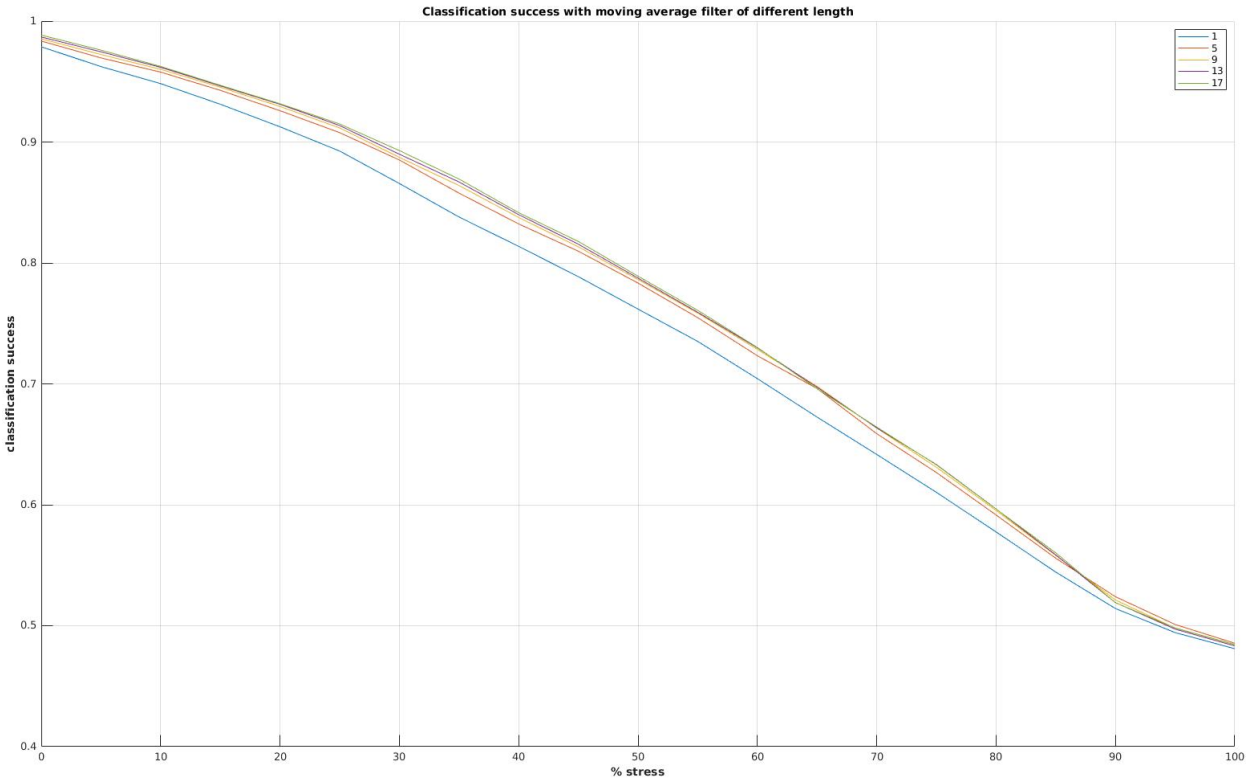

Figure 13. Classification success vs. image stress with stress types and video segments averaged. The family of curves are different-sized running averages.

Figure 14 shows a family of curves for different length running products when both the stress types and video segments are averaged. Figure 14 is very similar to Figure 13, but the running product filter shows a bit more increase of classification robustness than the running average filter.

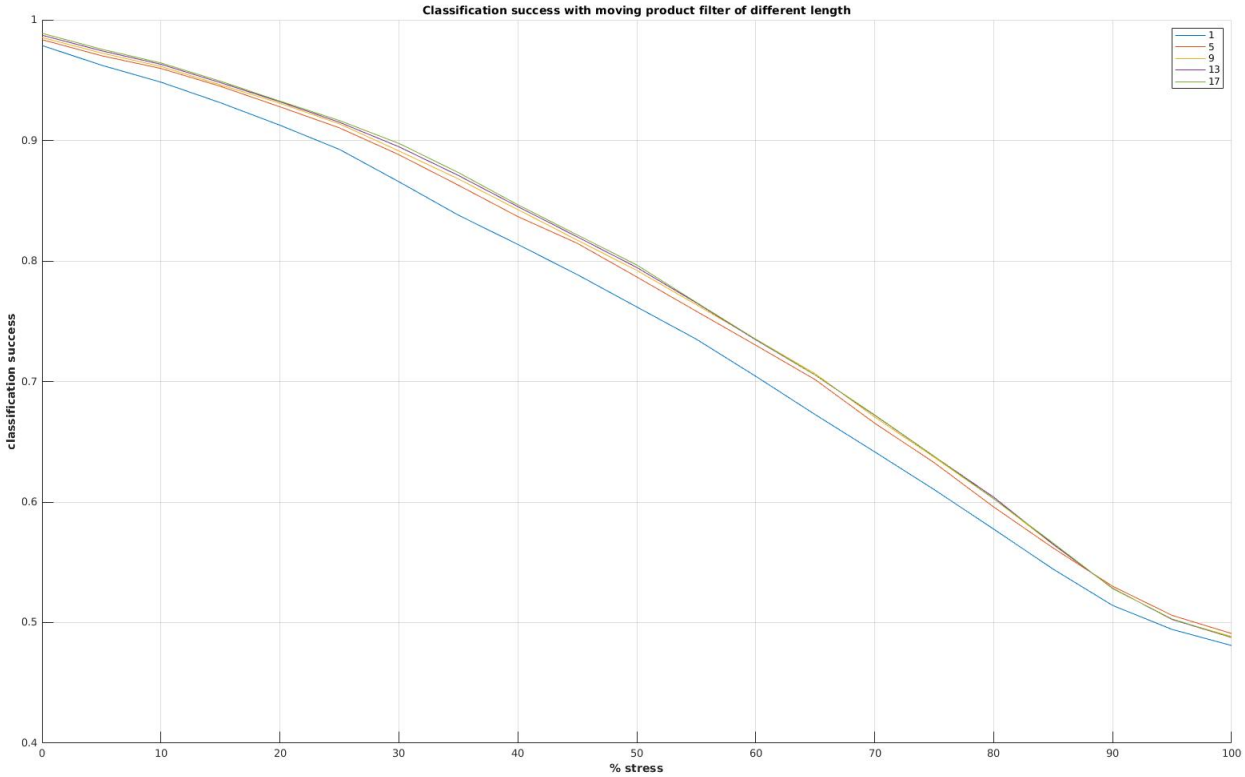

Figure 14. Classification success vs. image stress with stress types and video segments averaged. The family of curves are different-sized running products.

4. OBSERVATIONS AND CONCLUSIONS

There are a series of observations that must be made before drawing conclusions about this study. First, the semi-automated method that was used to create and identify the logical masks of the targets, the sky, and ground regions for many images had a lot of technical contortion and is far from being perfect. Approximately 20% to 30% of the resulting masked images had to be discarded because the regions and their labels were hard to stabilize automatically. Even the images that were kept were not always perfectly masked: some parts of the target were not included in the target mask, and the target mask occasionally included some of the background.

Because of these masking errors, the classification success curves do not drop to 0% success as the stress goes to 100%. This occurs because the stress is placed on the target mask, and some of the target stays unstressed for imperfect masks. In addition, stressing the background due to imperfect masks has no effect on the classification success of the classifiers.

Second, some of the sequences were too short to effectively apply the running average/running product filters. Even though the gain in robustness increases quickly after expanding the filter length to have more than one element in the running average/product, some of the shorter sequences clearly cannot have running averages of length comprising 20 elements. This may slightly affect the distribution of the family of curves shown in the plots.

Third, we could not control the change and rate of change of target aspect angle for every video segment. The video clips were downloaded from the web and thus were clips of opportunity. Although we tried to download clips with a slowly varying aspect angle, the triage needed to keep only the stable mask configurations sometimes scrambled the slow and gradual change of the aspect angle.

Finally, because of (1) imperfection in the masking and labeling of the target and background regions and (2) the lack of control over the change and rate of change of the aspect angle, no “quantitative” conclusions can be made from this study. However, these observations indicate that “qualitative” conclusions can be drawn. Figures 9–13 show that the use of a moving average/moving product filter on the class probabilities slightly increases the robustness of the classifiers. Moreover, the largest increase in classifier robustness is delivered with a relatively small-sized filter (three elements moving average/product). This is especially evident when looking at the plots of the classification success curves when both the stress type and video segments are averaged (Figures 13 and 14), which show a real, although modest, increase in robustness when using a running average/product filter.

The stress types most affected by introducing this method are camouflage, distortion, and hull defilade. The stress type moderately affected by this method is disruptive coloration, as implemented in this study. Finally, the stress type least affected by moving average filters is the jamming/obscuration stress, which affects the whole image instead of only the target image.

It is difficult to infer the validity of using this technique by comparing the different plots found in Figures 11 and 12 for different video segments because the segments show vehicles in different trajectories. Therefore, the change and the rate of change of the target aspect angle are not standardized between video segments and their increased robustness cannot be compared.

What the collected figures *do not* show is how this method, namely, the use of moving averages/product filters in adjacent video frames, can help classifier robustness in very heavily stresses battlefield environments. This is because most of the classification success plots still show relatively good classification ability when the stress completely covers the target mask. This in turn is due to the imperfect target masks that inadequately cover the target. A better masking and labeling method would be needed to completely cover the whole target so that the whole target could get stressed, which allows the classification success curves to go to 0% as the target stress goes to 100%.

These glimpses of increased classifier robustness should be reanalyzed using better controlled video clip data, instead of using data of opportunity found on the web. A careful study might prove that the use of a small moving product filter on the classification class probabilities would substantially increase classifier robustness in battlefield settings.

5. REFERENCES

1. Bhardarkar, S. (n.d.). *Assessing the limits of zero- and low-shot learning for real time decision* [unpublished manuscript] (support grant number W911NF-17-S-0003). School of Computing, University of Georgia.

2. Zeng, X., Wang, Z., & Hu, Y. (2022). *Enabling efficient deep convolutional neural network-based sensor fusion for autonomous driving*. ArXiv:2022.11231v1 [cs.CV].

3. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). *Large-scale video classification with convolutional neural networks*. 2014 IEEE Conference on Computer Vision and Pattern Recognition. p. 1725–1732. doi:10.1109/CVPR.2014.223.

4. Debroux, P. (2021). *(U) Analysis of artificial intelligence (AI) classification in stressed environments* (CCDC DAC-TR-2021-199). DEVCOM Data Analysis Center.

5. Debroux, P. S. (2022). *Analysis methodology of image classifiers in stressed environments* (DEVCOM DAC-TR-2022-089). DEVCOM Analysis Center.

LIST OF ACRONYMS

CNN	convolutional neural network
HMMWV	High Mobility Multipurpose Wheeled Vehicle
MRAP	Mine Resistant Ambush Protected

DISTRIBUTION LIST

DEVCOM Analysis Center
FCDD-DAE-E/P. Debroux
White Sands Missile Range, NM 88002

DEVCOM Analysis Center
FCDD-DAD-TP/E. Chatterton
Redstone Arsenal
Huntsville, AL 35898

DEVCOM Army Research Laboratory
FCDD-RLB-CI/Tech Library
2800 Powder Mill Rd.
Adelphi, MD 20783

Defense Technical Information Center
ATTN: DTIC-O
8725 John J. Kingman Rd.
Fort Belvoir, VA 22060-6218